# *Effective Teaching: What Is It and How Is It Measured?*

Steve Cantrell and Joe Scantlebury

*Robust, transparent feedback and evaluation systems are needed that recognize the inevitability of classification errors but work to reduce them as much as possible.*

At the heart of the student achievement gap lies a credibility gap. Our school systems are based on a premise we all know not to be true: that students are equally well served by whoever teaches their classes. The consequences – to students and to teachers – are great. The good news is that this open secret is no longer so; teachers and school leaders are talking about it and grappling with it. Few teachers now assert that teaching cannot be measured (Scholastic & Bill & Melinda Gates Foundation 2010). Design teams made up of courageous educators in numerous districts are engaged in the hard work of honestly rethinking their support and evaluation systems for teachers.

But amid this promise, there is also peril. If we're not careful about how we go about this work, we could replace one credibility gap with another. If teachers have reason not to trust the systems put into place to support and evaluate them, then these systems cannot achieve their aims of improving teaching effectiveness. If so, we will have lost a rare opportunity.

As states and school districts adopt systems to measure effective teaching, there is a growing concern about accuracy. Nobody wants a system that routinely misclassifies teachers. Some even assert that teaching cannot be measured: that teaching is an art, not a science, and dedicated teachers should not be subject to additional accountability pressures. But how do we balance those concerns with the needs of students? We cannot pretend that students are equally well served by whoever teaches them. Forgetting to balance students' concerns with those of teachers has dire consequences – ones that accrue disproportionately to young people already struggling to succeed.

Having the courage to walk this fault line between potentially misclassifying some teachers and not classifying teachers at all requires constant attention to the consequences for both teachers and students. It's a balancing act, to be sure; but if we cannot avoid error, we should err in favor of students. When building robust feedback and evaluation systems, perhaps it is best for us to admit that error is always present and be transparent about where it exists. In this way we build trust and limit misuse of feedback and evaluation systems.

## *Consequences for Students*

Findings from the teacher effectiveness literature reinforce what education professionals and those who have spent significant time in schools know well: the

*Steve Cantrell is senior program officer for research and evaluation and Joe Scantlebury is senior policy officer for U.S. Program Advocacy at the Bill & Melinda Gates Foundation.*

assignment of a student to a teacher's classroom is not a trivial exercise, but rather an act of great consequence.

This research literature can be reduced to three basic findings. Student performance differs across different classrooms, indicating that the quality of teaching matters (Rivkin, Hanushek & Kain 2005). Evidence from random assignment studies suggests that these differences are attributable to teachers, rather than to the student composition of the class (Kane & Staiger 2008). These differences are greater *within* schools than across schools, indicating that it is not enough to provide feedback and accountability at the school level (Nye, Konstantopoulos & Hedges 2004). Moreover, the performance differences are large. By some estimates, having a top quartile teacher versus a bottom quartile teacher yields performance gains equivalent to closing a quarter of the Black-White achievement gap (Gordon, Kane & Staiger 2006). In the Bill & Melinda Gates Foundation's Measures of Effective Teaching (MET) project,[1] these differences in student performance between those taught by top and bottom quartile teachers ranged from one-third to over a full year of learning gains. These are not minor differences.

Yet most school systems do little, if anything, to ensure that students have an equal chance to receive the best available instruction or to prevent students from being assigned to the least effective teachers for year after year. In many school systems, the status symbols and contractual arrangements work together to decrease the likelihood that students who struggle the most receive the most effective instruction. Too often, teacher status is determined by their students' performance level. Teachers of Advanced Placement, honors, or gifted students are accorded higher status than their peers whose students struggle in school. New teachers, who are demonstrably less effective than their more experienced peers, are not only given the last choice of assignment, but often have to teach multiple classes, each requiring separate preparation. These organizational features increase the difficulty of closing the achievement gap. In addition, the absence of robust measures of teaching effectiveness allows too many schools and districts to ignore these systemic inequities. While students, their parents and caregivers may not fully appreciate the magnitude of these systemic inequities, the impact on their lives is unmistakable.

Anecdotes are numerous of individual teachers who made a personal difference in a student's life. We are all familiar with these accounts. If we are regular readers of this journal, we can likely share stories of our own. Concluding that individual interventions

1 For more information on MET, see <www.gatesfoundation.org/united-states/Pages/measures-of-effective-teaching-fact-sheet.aspx>.

and instructional heroism is all that students and families can reasonably expect elevates these status privileges, contractual arrangements, and managerial omissions in ways that undermine the high aspirations of students, their families, and educators. Moreover, the absence of any clear or legislated right of students to an effective teacher creates no conflict of laws or balance of rights. Students have no enforceable right to an effective teacher, and thus they bear the burden of our systemic inequities.

Students have no enforceable right to an effective teacher, and thus they bear the burden of our systemic inequities.

### *Consequences for Teachers*

The most recent analyses fault teacher evaluation systems for their inability to differentiate among teachers (Weisberg et al. 2009). The typical system has two or three performance levels, yet assigns the lowest rating to less than one percent of all teachers. Teachers report that the evaluation process is often perfunctory. School leaders often receive minimal guidance and even less training on managing and executing teacher evaluation. When teachers have a positive experience with evaluation, it appears to be based on idiosyncratic factors, highly dependent upon the skills of the evaluator.

As a result, these weak feedback and evaluation systems are largely irrelevant to how schools conduct business. Seldom do feedback and evaluation systems inform consequential staffing and central office decisions. Even if those in charge know better, most school systems are organized as if differences among teachers were nonexistent. And when differences are apparent to teachers and patterns appear to disparately impact entire communities, school and district leaders seldom have the political courage or incentives to call the question about instructional practices. The measures in use seldom inform teacher assignment, professional development offerings, or promotion decisions.

The dire consequence for teachers is no feedback. Too many teachers are left alone to self-assess their competence and self-prescribe improvement. The difficultly of this bootstrapping effort is exacerbated by the relative isolation within which most teachers practice. The metaphor of the "egg-crate" school remains apt (Lortie 1975). Without accurate indicators and without meaningful exposure to other teachers' practice, self-improvement efforts are far from guaranteed to succeed. This is not mere conjecture: the data on returns to teacher experience shows little to no improvement beyond a teacher's fourth year of practice (Boyd et al. 2007). As Deborah Ball, dean of the University of Michigan School of Education (2011), said, "An enormous faith is placed on 'learning from experience,' despite substantial empirical evidence that experience is an unreliable 'teacher'" (p. 4).

The lack of any clear performance signal has other negative consequences for teachers, including uncertainty about whether they have satisfactorily accomplished their mission, a general disconnect between effort and reward, and growing unease with the system's failure to address teaching ineffectiveness (Rochkind et al. 2007). The lack of performance signals fails to encourage the right teachers to stay in the profession and the wrong ones to leave. While we certainly agree with Linda Darling-Hammond that "you can't fire your way to Finland" (UCLA/IDEA 2011),

we also believe that teachers come to the profession to do good and have hope that given stronger feedback, those few teachers who cannot succeed will leave teaching and find better ways to deploy their talents.

## *Increasing Trust*

Without trust, there cannot be feedback but only judgment. Only trustworthy information will be useful to teachers seeking to improve. Validity and reliability are the research standards for information quality and are useful ways to think about building trust in the information provided by feedback and evaluation systems. The Foundation's work with our MET partners has led us to focus on four "trustworthiness tests" – face validity, coherence, scoring reliability, and predictive validity.

*Face validity* is simply the "sniff" test. When teachers encounter the system for feedback and evaluation they want to see indicators that reflect competencies they value. To pass this test, teachers must believe that the system is directed toward aspects of teaching and learning that they believe make a difference to students. If the competencies required by the system could be met without fundamentally meeting the needs of students – "professional appearance" comes to mind – then teachers could attend to the competencies required by the system without influencing their ability to enhance student learning.

*Coherence* refers to the interconnections among parts of the system. If the feedback and evaluation system is unrelated or only loosely connected to other parts of the system that impact teaching and learning, such as professional development, curriculum and instruction, or mentoring, then opportunities for leveraging synergies across these areas are lost and the possibility increases for conflicting goals and confusion regarding outcomes. Importantly, the feedback and evaluation system should reflect the theory of instruction espoused by the district lest the disconnect between the two promotes confusion.

*Scoring reliability* – unreliability in scoring is the aspect of feedback and evaluation systems that may most undermine trust. Few school systems, however, routinely track or report rater reliability. For teachers (and their unions), it is patently unfair for their rating to be dependent upon the ability of the rater rather than the quality of the lesson. Our teacher advisory panel, our union partners, and the district administrators working closely with us all agree that uneven rater reliability is prevalent. In response to this need, we have plans to disseminate the training and monitoring methods used by the MET project researchers to ensure reliability.

*Predictive validity* indicates whether the system has the right focus. It refers to the association between competencies measured by the feedback and evaluation systems and the desired outcomes. If there is little or no association between the actions being tracked and the outcomes of value, then the system is broken. If this connection does not exist, then it is hard to support the claim that doing what the system requires will lead to the desired outcomes, such as increased student learning.

The MET project is an exercise in building trustworthy feedback and evaluation systems. It is not and never has been an attempt to build "the one best system." Instead, it serves to test the idea of a multi-faceted feedback and evaluation system by combining promising, yet emerging, indicators of teaching and learning. As MET serves to test an increasingly popular idea – multiple measures – it fully recognizes

## PERSPECTIVES: The Superintendent's View

*John Deasy is superintendent of the Los Angeles Unified School District.*

*What is the best way to address the objective of equitable access to high-quality instruction?*

To raise the performance levels of non-White, low-income students, parents in those communities need to be given viable educational options. Their children must no longer be forced to attend chronically underperforming schools. If another operator comes forward with a better plan to educate students in a low-income community, then it should be given the opportunity to do so. Only in this way can we begin to break the cycle of education failure that plagues too many of our students.

*From the administrators' point of view, are the performance management and instructional capacity-building strategies mutually exclusive? What else needs to be part of the discussion?*

Administrators and other practitioners must work closely with teachers to explain the meaning of teacher recommendations, particularly those that are based on new data and research. Teachers need to understand both strengths and weaknesses suggested by the data. In addition, teachers need to be made aware that Value Added and Academic Growth Over Time, among other measurements, are intended not to threaten their jobs, but to give them – plus parents and administrators – a better guide as to how they are doing their jobs.

We [also] focus on instructional capacity building strategies to improve student outcomes. Consequently, these strategies go hand in hand with performance management. The purpose of performance-based management is to ensure that an organization achieves its goals. As the superintendent, it is my responsibility to facilitate human performance that leads to improved student achievement. Performance management allows us to use data to determine in which instructional strategies to invest. As we continue to push for using data to foster accountability, we need to also use data to ensure that we truly become a learning organization.

that the promise of multiple measures is not that there are more measures, but that these measures represent different facets of teaching and learning that individually and collectively support student learning gains on outcome measures such as state performance assessments.

### *Reducing Error and Building Credibility*

There is a connection between reducing error, or misclassification, and increasing use. Teachers will use feedback only when they believe it will improve their practice. Otherwise, they will seek ways to game the system. Passing the "trustworthiness" tests goes a long way toward reducing error. Feedback is more likely to be used when the system is aligned with what teachers view as best practices; the parts of the system connect logically; scoring processes are reliable; and the indicators do, in fact, indicate what helps students learn better.

There are other types of error that similarly limit or distort the use of a feedback and evaluation system. Agreement around outcomes tops the list. When what is measured is disconnected from what is valued, efforts to increase scores on the measure will be met with little enthusiasm and even resistance. State assessments are routinely condemned as insufficient

(or even unfair) measures of school outcomes. There is reason for optimism on this front, as the consortia tasked with developing assessments aligned to the Common Core Standards will likely improve the substance and status of state tests. Still, it would be too easy to use the need to improve tests as a reason to avoid accountability and feedback – if the outcome is important to student success, measure it.

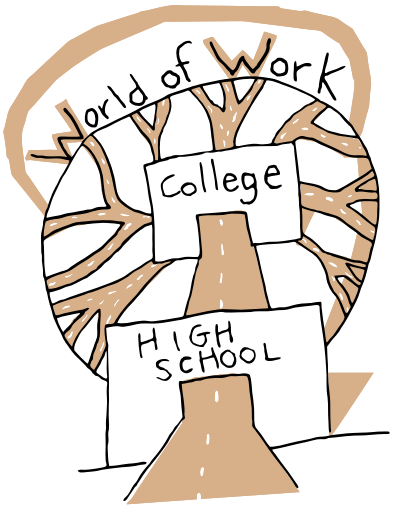
Attribution is a thorny problem that, left unresolved, also will undermine the feedback and evaluation system. At the most basic level, there is the administrative challenge of ensuring that the data systems link the right students to the right teachers. This sounds deceptively simple, yet it is quite common for the teacher of record to be different from the teacher who provided the instruction. In many elementary schools, students are re-grouped for math and/or English language arts. While the school may know perfectly well which students are taught by which teachers and for what duration, the central office records may not be accurate. It is easy to see the damage to the system's credibility should a teacher receive feedback (or be rewarded or sanctioned) based on students taught by another teacher.

Related to the attribution problem is where to place accountability. Accountability for effective teaching cannot sit solely upon the shoulders of teachers. If supports are deployed, as a school system seeks to close the gap between the most and least effective teachers, then the effectiveness of these supports should be subject to the same rigorous feedback and evaluation processes. If a particular professional development or curricular intervention does not improve performance for those who have received it, then the system cannot claim to have supported teacher development. Similarly, if the working conditions at a school do not increase the likelihood that those teachers who struggle are supported by their more successful colleagues, then the administration of that school is failing to support teacher growth and needs assistance. The fact that measures are precise at the teacher level does not limit their use to that level.

Finally, we return to misclassification. So far, researchers have not been able to explain what appears to be an anomaly in the empirical findings – persistent and consequential differences in student performance for top and bottom quartile teachers alongside apparently unstable teacher rankings. It appears inconsistent to hold both findings as true. If teachers are routinely misclassified, why, when compared to similar groups of students, do the students of previously identified top and bottom quartile teachers persistently outperform (for top quartile teachers) or underperform (for bottom quartile teachers) their peers?

We can only speculate why misclassification exists: it could be that a majority of teachers provide similar instruction and only the top and bottom 15 percent meaningfully differ from the average; or even the top and bottom 5 percent or 10 percent. We don't know. It matters because many of the state and district evaluation systems assume that it is possible to accurately assign teachers to one of three or four rating categories.

To build trust means not eliminating error, but committing to reduce it. We can reduce the error of misclassification if we focus on where we think we have the best information. If not, again, we could replace one credibility gap with another – pretending that teachers fall neatly into four or more




categories of effectiveness – when we do not know how many categories exist or whether our measures are good enough to make such fine distinctions.[2]

The MET project will explore this anomaly in an upcoming report based on over 12,000 lessons captured on video. The analysis of teacher practice will provide an estimate of observable differences among teachers and provide some evidence to suggest how large the "messy middle" of teacher practice is.

2 One path forward is to increase our understanding of the true performance distribution – it's not likely normal. The size of the middle part of the distribution matters. A purely hypothetical example will help illustrate the point. Assume that 70 percent of teachers constitute a middle where it is difficult to find observable differences in teaching practice. In this case, the underlying distribution of teacher practice would be 15 percent observably weaker than average, 70 percent average, and 15 percent observably stronger than average. If the categories used to differentiate teaching quality do not reflect the underlying distribution, but used quartiles instead, the misclassification rate is by definition at least 40 percent at both the highest and lowest quartiles. Moreover, since these teachers' practice is indistinguishable from average practice, those misclassified at either the top or bottom quartile could be categorized in the opposite quartile the following year. While 40 percent would indicate an unacceptable level of misclassification, if the remaining 60 percent of teachers in each of these quartiles were identified correctly (the real top and bottom performers), large performance differences between students of top and bottom quartile teachers would persist from year to year.

## *Implications for Civil Rights*

Most Americans share the value that all students deserve an equal opportunity to receive a high-quality education. We understand that individual student effort and motivation, coupled with family and community support and expectations, may play a part in the success of an individual student. We also understand that even without those supports, students can graduate ready for college and careers, if they have teachers dedicated to this mission. Thus, an equal opportunity to a high-quality education should, at minimum, afford every child a chance to be taught by the best teachers that a school system has to offer. If for some reason whole groups of students were denied this chance, or if the opportunity to be taught by a great teacher were nothing more than chance, we would collectively demand that such a system be changed.

The scenario is not hypothetical. We know that many students are routinely provided with the least effective instruction. This directly impacts and perpetuates the so-called academic achievement gap – a gap that W.E.B. Du Bois (1903) wrote about eloquently in *The Souls of Black Folk*. In this seminal work, Dubois described education's potential to lift a people newly emancipated and striving to overcome the pernicious effects of Jim Crow laws and stark racism. He observed that education was essential both for sustenance and citizenship and hoped that "Education [would] set this tangle straight" (p. 91). He charged educators at the turn of the last century to embrace that mission and unflaggingly prepare the next generation.

Du Bois would be pleased to know that such educators exist among the current generation. As we work in partnership with teachers to determine

what it means to be effective, we are increasingly aware that current teachers are not monolithic in their views, or blind to the deleterious impact on students of teacher assignment, distribution, evaluation, and support practices that relegate the neediest students to instructional settings with the least potential for success. These teachers, conscious of the classroom and life challenges that students face, seek ways to support and spread great teaching practices, improve instruction, and fairly transition out of the profession colleagues for whom it is not a good fit. We support and seek to inform their efforts. Together, we are clear that closing achievement gaps will not happen by chance or by avoiding serious conversations about what we owe students, whose uncodified rights do not include the right to an effective teacher.

While it may not be a right, fairness dictates that school systems at the very least know which of its students receive instruction from the least effective teachers and take measures to ensure that this doesn't happen to particular students year after year. In the longer run, closing the teaching effectiveness gap – and thereby reducing the consequences accompanying assignment to the least effective teachers – is perhaps the single most important step we can take toward closing the achievement gap. This requires measures that we can trust, so that systems know which teachers are most in need of support and which students, having suffered inadequate instruction, require special handling to ensure that this does not happen in consecutive years. Most importantly, these measures should provide trustworthy feedback. For it is through feedback that we get to Finland. The path to improvement cannot possibly lead through ignorance.

*References*


Boyd, D. J., H. Lankford, S. Loeb, J. E. Rockoff, and J. H. Wyckoff. 2007. *The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools*. CALDER Working Paper 10. Washington, DC: The Urban Institute.

Du Bois, W.E.B. 1903. *The Souls of Black Folk*. Chicago: A. C. McClurg & Co.; Cambridge, MA: University Press John Wilson and Son.

Gordon, R., T. J. Kane, and D. O. Staiger. 2006. *Identifying Effective Teachers Using Performance on the Job*. Discussion Paper 2006–01. Washington, DC: Brookings Institution, Hamilton Project.

Kane, T. J., and D. O. Staiger. 2008. *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*. National Bureau of Economic Research Working Paper 14607 (December). Cambridge, MA: NBER. Available for download at <www.nber.org/papers/w14607>.

Lortie, D. 1975. *Schoolteacher: A Sociological Study*. Chicago, IL: University of Chicago Press.

Nye, B., S. Konstantopoulos, and L. V. Hedges. 2004. "How Large are Teacher Effects?" *Educational Evaluation and Policy Analysis* 26:237–257.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. "Teachers, Schools, And Academic Achievement," *Econometrica* 73, no. 2 (March):417–458.

Rochkind, J., A. Ott, J. Immerwahr, J. Doble, and J. Johnson. 2007. *They're Not Little Kids Anymore: The Special Challenges of New Teachers in High Schools and Middle Schools*. Lessons Learned: New Teachers Talk about Their Jobs, Challenges and Long-Range Plans, Issue no. 1. New York: National Comprehensive Center for Teacher Quality and Public Agenda.

Scholastic and the Bill & Melinda Gates Foundation. 2010. *Primary Sources: America's Teachers on America's Schools*. New York: Scholastic. Available for download at <www.scholastic.com/primarysources>.

UCLA/IDEA. 2011. "An Education Exchange with Linda Darling-Hammond," UCLA/IDEA Newsroom (April 1), <http://idea.gseis.ucla.edu/newsroom/idea-news/an-education-exchange-with-linda-darling-hammond>.

University of Michigan School of Education. 2011. "The Whole Is Greater Than the Sum of the Parts," *Innovator* 41, no. 1 (Spring). Available for download at <www.soe.umich.edu/information_for/alumni_friends/innovator_magazine/volume_41_spring_2011>

Weisberg, D., S. Sexton, J. Mulhern, and D. Keeling. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. New York: New Teacher Project.